Rethinking Reprojection:

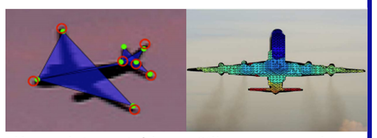
Closing the Loop for Pose-aware Shape Reconstruction from a Single Image

Rui Zhu, Hamed Kiani Galoogahi, Chaoyang Wang, Simon Lucey
The Robotics Institute, Carnegie Mellon University

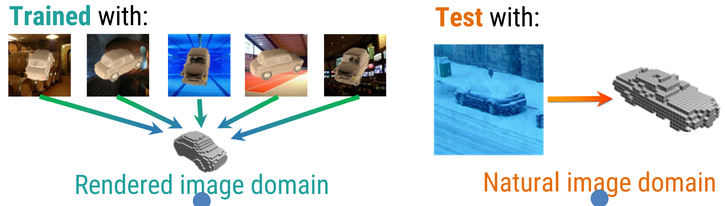
ICCV 17
International Conference on Computer Vision 2017

1. Motivation

- **Traditional methods** purely base on geometric cues of **reprojection error**.



- **Deep methods** uses direct regression in a data-driven way:

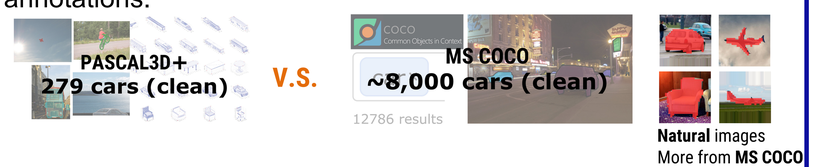


- **Problems**:

1. **Poseless** → Cannot measure silhouette matching;
2. Cannot train on natural images because annotating shape & pose on natural images are expensive; training with rendered images instead. → **'Feature Gap"** → Poor generalization.

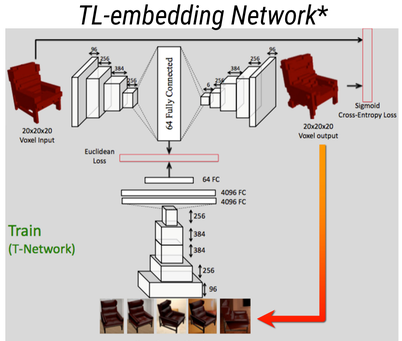
- **Observation**:

Abundant annotation of instance segmentation masks than 3D annotations are readily available. We should utilize these 2D annotations.



By chaining the output shape back to the image, we can train (finetune) with **weak supervision** of **silhouette reprojection error**, on the target **natural image** domain.

*TL-embedding Network**



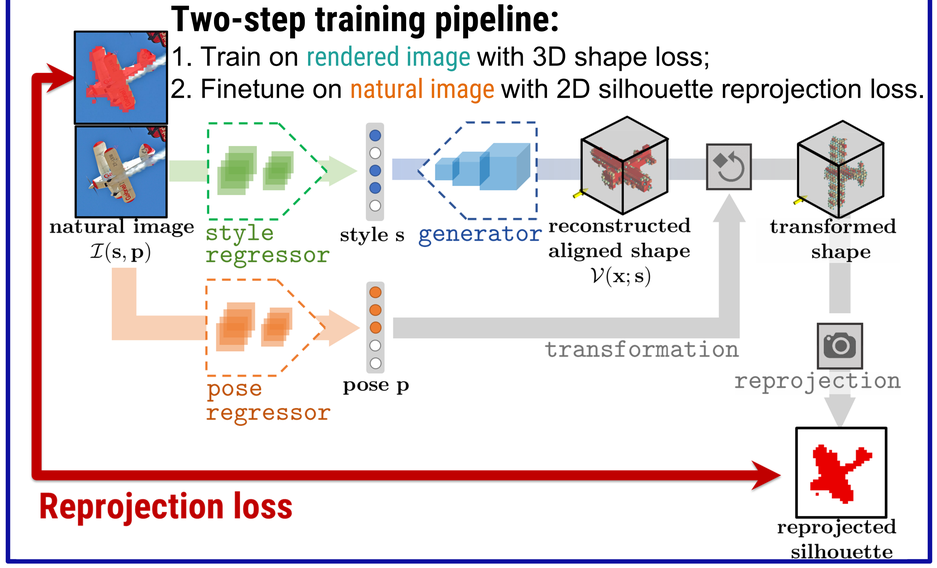
Reprojected silhouette (orange) should match with the annotated mask (grey)

* *Girdhar, Rohit, et al. "Learning a predictable and generative vector representation for objects." European Conference on Computer Vision. Springer International Publishing, 2016.*

2. Method

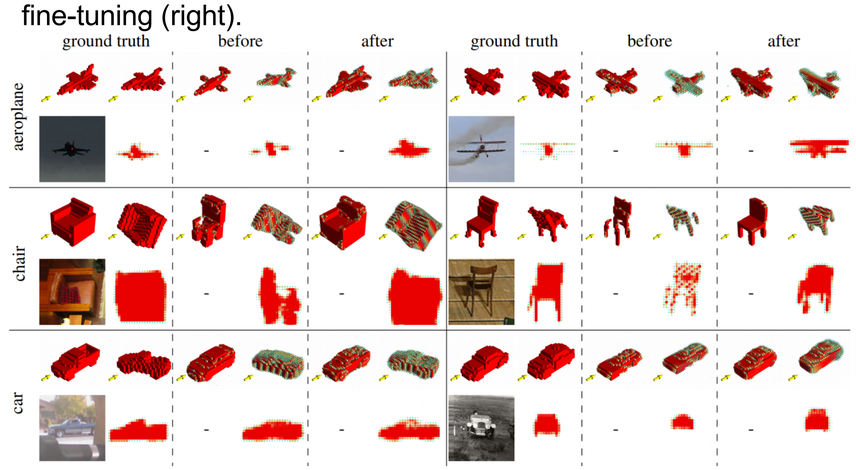
Two-step training pipeline:

1. Train on rendered image with 3D shape loss;
2. Finetune on natural image with 2D silhouette reprojection loss.

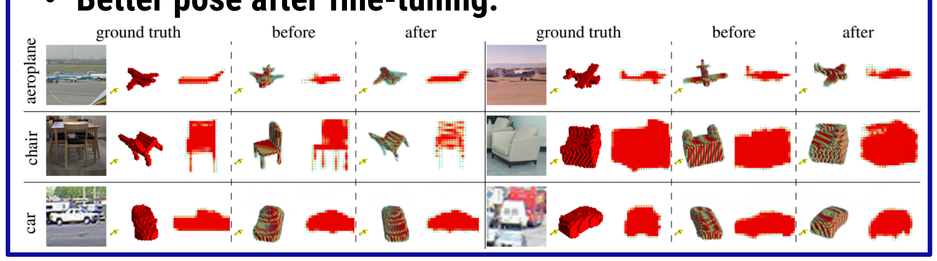


4. More Qualitative Results

- **Better style after fine-tuning.** For each sample, illustrations include: input image, aligned & shape-aware shapes, reprojected silhouette of ground truth (left), before fine-tuning (middle) and after fine-tuning (right).

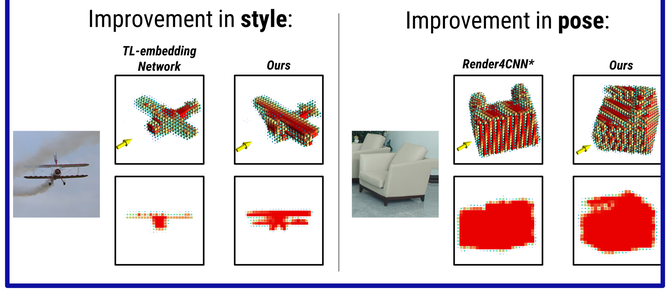


- **Better pose after fine-tuning.**



3. Results in a Nutshell

Improved style & pose estimation on natural images after fine-tuning with reprojection loss on natural images.



5. Quantitive Results

Datasets:

- PASCAL 3D+ with ground truth in shape & pose;
- MS COCO with ground truth in masks.

	aeroplane	chair	car
rendered with shapes	206,296	345,001	382,144
MS COCO with masks	4,734	3,200	2,942
PASCAL3D+ with shapes	125	220	279

Two Models:

- p-TL uses encoder-decoder,
- p-3D-VAE-GAN uses VAE-GAN.

Comparisons:

- before (in the table) is training only on rendered images;
- after (in the table) is after finetuning with our method.

Measuring:

- 2D AP against annotated silhouettes, and 3D AP against annotated shapes;
- Error in pose estimation.

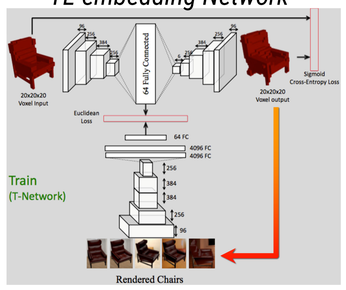
		p-TL			p-3D-VAE-GAN		
		aero	chair	car	aero	chair	car
2D AP	before	0.589	0.844	0.815	0.627	0.852	0.851
	after	**0.704**	**0.849**	**0.872**	**0.720**	**0.878**	**0.894**
3D AP	before	0.211	0.531	0.630	0.183	0.527	0.642
	after	**0.219**	**0.552**	**0.639**	**0.249**	**0.577**	**0.664**
rotation $Acc_{\frac{\pi}{6}}$ / $MedErr$	before	0.67/23.0	0.78/**8.2**	**0.83/4.8**	0.67/23.2	0.76/8.2	0.86/5.0
	after	**0.68/17.3**	**0.80**/8.3	0.80/5.2	**0.70/17.2**	**0.80/8.1**	0.86/**4.7**
	Su et al.	0.76/15.1	0.85/9.7	0.86/6.1	0.76/15.1	0.85/9.7	0.86/6.1
translation $MedErr$	before	0.092	0.074	0.060	0.088	0.079	0.061
	after	**0.077**	**0.072**	**0.058**	**0.073**	0.079	**0.050**

ICCV17
International Conference on
Computer Vision 2017

2. Main Idea

- Chain the output shape back to the image, so that we can train (finetune) with weak supervision of silhouette reprojection error, on the target natural image domain.

*TL-embedding Network**



Reprojected silhouettes should match with the annotated mask

ICCV17
International Conference on
Computer Vision 2017

ICCV 17
International Conference on
Computer Vision 2017